REVISTA ESPAÑOLA DE DOCUMENTACIÓN CIENTÍFICA
31, 3, JULIO-SEPTIEMBRE, 327-343, 2008
ISSN 0210-0614

ESTUDIOS / *RESEARCH STUDIES*

Solapamiento y singularidad de MEDLINE, WoS e IME para el análisis de la actividad científica de una región en Ciencias de la Salud

Overlapping and singularity of MEDLINE, WoS and IME for the analysis of the scientific activity of a region in Health Sciences

Rodrigo Costas*, Luz Moreno*, María Bordons*

Resumen: La inclusión de un solo lugar de trabajo en los registros bibliográficos de MEDLINE constituye una conocida e importante limitación para el uso de la base de datos con fines bibliométricos. En este documento se ofrecen datos cuantitativos sobre las repercusiones que esta limitación tiene sobre el volumen e impacto de la producción científica de una determinada región española en un estudio basado en MEDLINE. En concreto, se estima que 1/3 de los documentos de la región no se identifican en esta base de datos por figurar la región en una posición distinta a la primera; y tiende a infravalorarse el impacto regional, ya que son documentos en colaboración que se publican en «mejores» revistas y reciben más citas que el promedio de la producción.

Palabras clave: ciencias de la salud, bases de datos bibliográficas, lugar de trabajo, solapamiento, cobertura, singularidad, bibliometría, MEDLINE, Web of Science, IME.

Abstract: The inclusion of a single affiliation address in MEDLINE bibliographic records is a well-known important limitation for the use of this database for bibliometric purposes. This paper provides quantitative data regarding the effects of this limitation on the quantity and impact of the scientific production of a Spanish region measured via MEDLINE. It is estimated that 1/3 of the papers produced by this region cannot be identified in this database because the region does not appear in the affiliation address. Moreover, the impact of the region tends to be undervalued since

* Instituto de Estudios Documentales sobre Ciencia y Tecnología (IEDCYT, antes CINDOC). CSIC. Madrid (España). Correo-e: rodrigo.costas@cindoc.csic.es.
Recibido: 11-10-2007; 2.ª versión: 30-1-08.

these collaborative papers are published in «better» journals and receive more citations than many other publications.

Keywords: health sciences, bibliographic databases, institutional address, overlapping, coverage, singularity, Bibliometrics, MEDLINE, Web of Science, IME.

1. Introducción

Las bases de datos bibliográficas desempeñan un papel fundamental en la investigación bibliométrica, ya que permiten analizar la actividad científica realizada por investigadores, centros, regiones y países; detectar sus fortalezas y debilidades e identificar tendencias en la investigación. Las bases de datos multidisciplinares –por ejemplo, el Web of Science o, más recientemente, Scopus– son muy adecuadas para el estudio de la actividad de un país o región en todas las áreas del conocimiento, mientras que las bases de datos especializadas adquieren especial relevancia en los estudios de áreas temáticas concretas, ya que, en general, presentan una mayor cobertura del área y un tratamiento más elaborado de la información temática (descriptores, términos de indización, códigos de clasificación, etc).

La selección de las bases de datos a utilizar en un estudio bibliométrico constituye una fase esencial del mismo, ya que existen diferencias entre ellas en su cobertura, en la información proporcionada para cada registro y en sus prestaciones de descarga de documentos. En lo que se refiere al vaciado de fuentes primarias, algunas bases de datos vacían solo artículos de revistas, mientras que otras incluyen libros, actas de congresos, e incluso patentes o informes. Por otro lado, algunas bases sólo proporcionan la información bibliográfica básica de cada registro (autor, título, revista, etc.), frente a otras que añaden el lugar de trabajo de los autores, o indizan los documentos (palabras clave, códigos de clasificación). Especialmente relevante es la exhaustividad y grado de normalización de la información contenida en algunos campos como el campo «autor» (Costas y Bordons, 2007; Ruiz-Pérez et al., 2002; Spinak, 1995) o el «lugar de trabajo» de los investigadores (Gálvez y Moya-Anegón, 2007; Gómez y Galban, 1986). Todas estas características determinan la selección de una u otra fuente y, en última instancia, repercuten sobre los resultados de los estudios.

En el área de las Ciencias biomédicas, MEDLINE es sin duda la base de datos de referencia por su amplia cobertura temática y la indización de sus documentos a través de un lenguaje controlado. Pero existen algunas limitaciones en su uso bibliométrico, como es la poca normalización de algunos de sus campos (autores, centros, países) y la inclusión de un solo lugar de trabajo. Este último factor limita su uso como base de datos única en los estudios bibliométricos en que se emplea una delimitación geográfica, es decir, aquellos en los que se plantea una búsqueda basada en el campo lugar de trabajo, como es el caso de la identificación de los documentos

de una región o un país, ya que sólo se recuperan aquéllos en los que la región o país estudiado aparece como primer firmante. Aunque esta limitación es bien conocida, no se ha descrito cómo puede repercutir en los resultados de los estudios. En concreto, nos planteamos en este trabajo cuál es la recuperación que ofrece MEDLINE, frente a la ofrecida por el Web of Science (WoS) o el Indice Médico Español (IME), de la producción científica de una determinada región española en Ciencias de la Salud, qué porcentaje de su producción no se recupera por incluir MEDLINE un solo lugar de trabajo, y qué repercusiones puede tener esta pérdida de documentos sobre el impacto de la región obtenido en los estudios. Dado que la producción no recuperada comprende documentos en colaboración (todos tienen más de 1 lugar de trabajo), para los que habitualmente se describe un mayor impacto esperado y observado (Bordons et al, 1993; Torres Salinas, 2007), ¿podría infravalorarse el impacto medio de la región a causa de esta pérdida de documentos?

Recientemente se ha realizado en el IEDCYT un análisis de la producción científica de la Comunidad de Aragón en Ciencias de la Salud (Gómez et al., 2007b). En dicho estudio se utilizaron tres bases de datos para obtener una cobertura lo más amplia posible del área: a) MEDLINE, principal base de datos en Ciencias de la Salud, que presenta una amplia cobertura de revistas biomédicas. b) WoS, base de datos multidisciplinar e internacional, que recoge una selección de revistas de alta calidad y prestigio, y aporta indicadores específicos como el factor de impacto de las revistas. c) IME, base de datos española especializada en Medicina, que aporta a las bases anteriores la investigación de interés más local. En este estudio se analizan los aspectos antes mencionados tomando como unidad de análisis la producción científica en Ciencias de la Salud de esta región.

2. Objetivos

Se plantea el estudio de la producción científica de Aragón durante 2001-2005 en la base de datos MEDLINE, así como en WoS e IME, con el fin de analizar los siguientes aspectos:

- solapamiento y singularidad de las bases de datos MEDLINE, WoS e IME en su cobertura de las publicaciones de la región;
- cuantificar el volumen de documentos no recuperados de MEDLINE por no pertenecer a esta región el primer centro firmante de las publicaciones (único recogido en esta base de datos);
- analizar la posible repercusión que la recuperación incompleta de documentos puede tener sobre las medidas más habituales de impacto de una región.

Se desean extraer conclusiones útiles de cara a la realización de futuros estudios

bibliométricos en el área de Ciencias de la Salud. Aunque los datos aquí analizados se refieren a una determinada región española, ésta es una región de tamaño medio cuyo comportamiento podría ser representativo del de otras comunidades españolas, aunque confirmar este último aspecto queda fuera de los objetivos de este trabajo.

3. Metodología

3.1. Descripción de las Bases de Datos: MEDLINE, WoS, IME

MEDLINE (http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html)

Base de datos bibliográfica elaborada por la National Library of Medicine (NLM) de Estados Unidos. Presenta una cobertura temática amplia en biomedicina y salud, abarcando también ciencias de la vida, del comportamiento, ciencias químicas y bioingeniería; todas ellas son disciplinas necesarias para los profesionales de la salud, y para otros profesionales dedicados a la investigación básica en ciencias biomédicas, a la atención sanitaria, a la sanidad pública y al desarrollo de políticas sanitarias. Incluye documentos publicados en más de 5.000 revistas y en 37 idiomas diferentes. El 90% de las revistas se recogen «cover to cover». Un aspecto distintivo de MEDLINE es que los registros son indizados con el vocabulario controlado de la NLM, los Medical Subject Headings (MeSH), lo que facilita las búsquedas de información. En los primeros años, para cada documento se incluía un máximo de seis autores, añadiendo luego «et al» para indicar la presencia de otros colaboradores. Desde el año 2000 no existe límite en el número de autores, y desde 2002 aparecen los nombres completos de los mismos. En los primeros años tampoco se recogía el lugar de trabajo de los autores, y a partir de 1988 se incluye la dirección institucional del primer autor del artículo. El campo de la dirección del autor no está normalizado, apareciendo la dirección tal como es especificada por el autor, pudiéndose encontrar en inglés y/o en español. En algunos registros no aparece el país, sino la provincia, la capital de provincia o la ciudad y en otros incluso no aparece el campo dirección (Rodríguez Gairín y Somoza Fernández, 2003). La presencia de un solo lugar de trabajo y la falta de normalización complica el uso de la base de datos con fines bibliométricos. En este estudio se empleó la base de datos MEDLINE disponible a través del Web of Knowledge.

Web of Science (WoS) (http://scientific.thomson.com/products/wos/)

Elaborada por Thomson Scientific (antes Institute for Scientific Information, ISI), comprende tres importantes bases de datos: Science Citation Index (SCI), Social Sciences Citation Index (SSCI) y Arts & Humanities Citation Index (A&HCI). Es una base de datos bibliográfica multidisciplinar e internacional que proporciona información sobre aproximadamente 9.300 revistas científicas de gran

prestigio internacional. El WoS presenta diversas características que la convierten en la base de datos más utilizada con fines bibliométricos: recoge todos los autores de un documento y la dirección de todos ellos –con cierto grado de normalización–, lo que permite realizar estudios de colaboración científica; y además incluye las referencias de cada documento, así como el número de citas que recibe, siendo posible mediante esta información calcular distintos indicadores basados en citas. Por otro lado, Thomson publica anualmente el Journal Citation Reports (JCR), que incluye una elaborada recopilación de indicadores sobre las revistas incluidas en el SCI, entre ellos el factor de impacto de las revistas, útiles para identificar las revistas más prestigiosas dentro de cada disciplina. El WoS presenta cierto sesgo a favor de la ciencia básica frente a la aplicada y a favor de las revistas en lengua inglesa (Moed, 2005).

Indice Médico Español (IME) (http://bddoc.csic.es:8080)

Esta base de datos recoge en la actualidad una selección de más de 200 revistas médicas españolas. Cubre todas las disciplinas básicas, las especialidades clínicas y las áreas relacionadas con aspectos asistenciales, organizativos y metodológicos de la medicina y campos afines. Para cada registro se incluye un conjunto de términos y códigos que representan el contenido del documento (descriptores, palabras clave). La base de datos recoge el lugar de trabajo de todos los autores, pero en los últimos años, debido a una serie de vicisitudes por las que atraviesa esta base datos, este campo falta en parte de los documentos (Osca Lluch, 1999).

3.2. Delimitación y obtención de la población de estudio.

Se ha obtenido la producción de Aragón en las bases de datos MEDLINE, WoS e IME durante el periodo 2001-2005 (Gómez et al., 2007b) a través de una búsqueda en el campo lugar de trabajo. Dicha búsqueda incluyó el nombre de las provincias, así como el de las principales ciudades de la región, esto último necesario en MEDLINE porque este campo no está normalizado. Todos los documentos de la región incluidos en las bases de datos MEDLINE e IME, especializadas en medicina, se consideraron relevantes para el estudio. En la base de datos WoS, la delimitación temática se hizo atendiendo a la clasificación de revistas en disciplinas elaborada por Thomson Scientific; de forma que se consideraron documentos médicos aquéllos publicados en alguna de las revistas incluidas en las disciplinas que se presentan en el Anexo I.

Los documentos obtenidos a partir de las diferentes bases de datos se enfrentaron entre sí a través de distintos algoritmos, que se basan en la comparación iterativa de los registros de las distintas bases de datos a través de diferentes campos (título del documento, título de revista, fecha publicación, etc.) lo que permite detectar

documentos solapados en diferentes pasos y con alto grado de acierto (Costas e Iribarren-Maestro, 2007).

3.3. Indicadores de solapamiento

Con el fin de cuantificar el solapamiento de documentos entre las diferentes bases de datos, se han utilizado las siguientes medidas:

– Índice de Meyer (Meyer et al., 1983), también denominado índice relativo de singularidad o peculiaridad (Cañedo Andalia, 1999; Pulgarín y Escalona, 2007). Este indicador permite comparar la cobertura de varias bases de datos sobre un tema determinado. Para este índice, las fuentes primarias únicas, contenidas en una sola base de datos, son las que tienen un mayor peso o valor (peso=1), que se reduce progresivamente para las fuentes duplicadas (peso=0,5) o triplicadas (peso=0,3). De este modo se premia las bases de datos que presentan más documentos de forma única. Cuánto mayor es el índice, mayor es la singularidad de la base de datos, es decir, que ésta recoge un mayor número de documentos únicos, sólo recogidos en dicha base, lo cual tiene un gran interés para realizar una adecuada selección de fuentes de información en los estudios bibliométricos.

$$\text{Índice Meyer} = \frac{\text{Sumatorio número de documentos * Peso}}{\text{Núm. total documentos recuperados}}$$

– % solapamiento relativo. Es una medida usada originalmente por Bearman y Kunberger (1977) y definida por Gluck (1990). Consiste en calcular el solapamiento de una base de datos en otra, teniendo en cuenta el peso de los documentos solapados respecto de los de presencia única. La fórmula general es:

$$\%\ \text{solapamiento en A} = \frac{|A \cap B|}{|A|}$$

$$\%\ \text{solapamiento en B} = \frac{|A \cap B|}{|B|}$$

3.4. Indicadores bibliométricos

Para analizar el impacto de la producción de la región se han obtenido los siguientes indicadores:

– Factor de impacto anual de las revistas de publicación de los documentos. A cada documento se le ha asignado el Factor de Impacto (FI) de su revista de publicación en el año correspondiente (Journal Citation Report, JCR). El Factor de Impacto es un indicador del prestigio de las revistas científicas. Puede considerarse un indicador del «impacto esperado» de un documento.
– Número total de citas. Total de citas recibidas por cada documento desde su año de publicación hasta la actualidad. Es un indicador de la visibilidad o «impacto observado» de cada documento en la comunidad científica.

Dado que sólo se dispone de datos de impacto para los documentos contenidos en WoS, este análisis se centra en el conjunto de documentos publicados en revistas comunes a WoS y MEDLINE. Se analiza si hay diferencias en el impacto de los documentos firmados por la región en primera posición y aquéllos firmados en posteriores posiciones (estos últimos no identificados por la búsqueda en MEDLINE, pero sí en WoS).

Se utilizan tests para muestras no paramétricas y la U de Mann-Whitney, considerándose las diferencias significativas cuando $p < 0,05$.

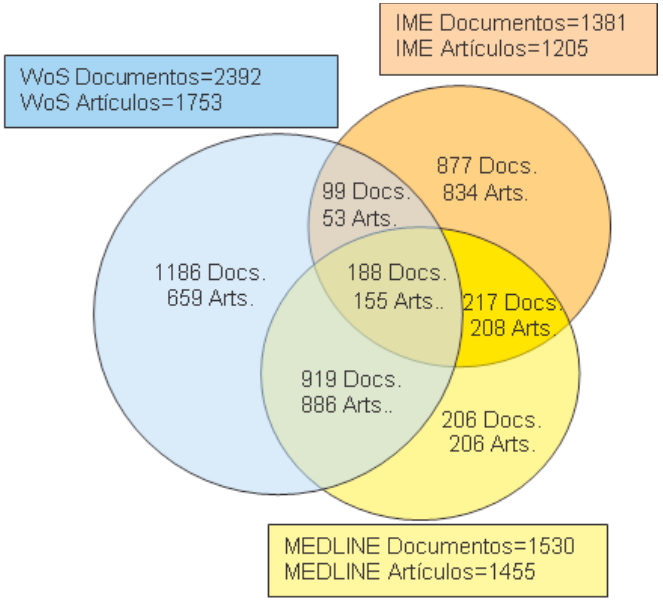
4. Resultados

4.1. Cobertura y solapamiento de documentos entre bases de datos

Se identifican 1.530 documentos en MEDLINE, frente a 2.392 en WoS y 1.381 en IME. Las tres bases de datos analizadas proporcionaron un total de 3.692 documentos diferentes, publicados en un total de 1.114 revistas. En total, 2.269 documentos (62%) son documentos únicos, recogidos en una sola de las bases de datos, y 1423 (38%) están solapados entre dos o más bases de datos. En la figura 1 se presenta la distribución del total de documentos y artículos por bases de datos, así como los datos de solapamiento entre ellas.

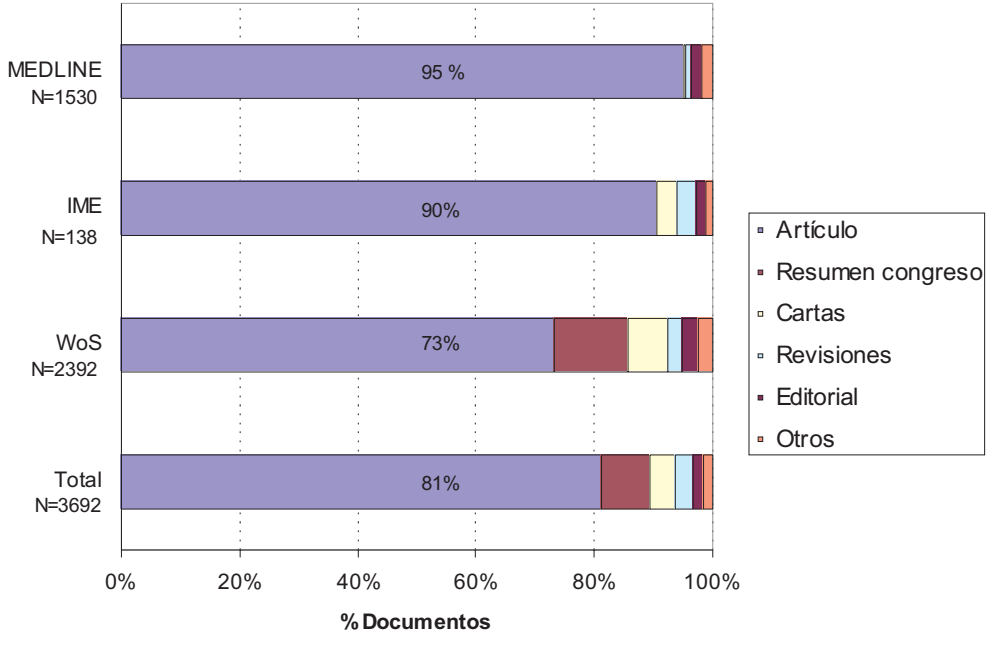
En las tres bases de datos predominan los artículos de revista, pero su proporción varía desde el 95% de artículos en MEDLINE hasta el 75% en WoS. En esta última base de datos se incluye un mayor porcentaje de resúmenes de congresos y cartas. Las distintas políticas de indización de las bases de datos, que difieren en su cobertura de los tipos documentales, contribuyen a explicar estas diferencias y deben tenerse en cuenta en los estudios de solapamiento ya que algunas revistas pueden estar incluidas en dos bases de datos, pero no todos sus documentos están duplicados. Los análisis incluidos en este documento se limitan al tipo documental «Artículo», por ser el que tiene mayor importancia en la transmisión de resultados originales de investigación.

Figura 1
Datos generales del solapamiento de documentos entre bases de datos



Nota: Total = 3.692 documentos; 3.001 artículos.

Figura 2
Tipología documental por bases de datos



4.1.1. «Singularidad» de las bases de datos

El análisis de la «singularidad» de las diferentes bases de datos se puede realizar a través del porcentaje de documentos «únicos» a cada base de datos, pero también a través del índice de Meyer, que no sólo considera los documentos únicos, sino el grado de solapamiento con las otras bases. A través de ambos indicadores se observa la mayor singularidad de IME, seguida de WoS. MEDLINE presenta la menor singularidad, debido principalmente a su gran solapamiento con WoS (tabla I).

Tabla I
Singularidad de las distintas bases de datos (artículos)

Base de datos	*% Documentos únicos*	*Índice Meyer*
WoS	659*100/1.753=38%	$\frac{659*1+939*0,5+155*0,3}{1.753} = 0,67$
MEDLINE	206*100/1.455=14%	$\frac{206*1+1.094*0,5+155*0,3}{1.455} = 0,55$
IME	834*100/1.205=69%	$\frac{834*1+263*0,5+155*0,3}{1.250} = 0,81$

4.1.2. Análisis del solapamiento

El porcentaje de solapamiento de cada base de datos con respecto a cada una de las otras dos se muestra en la tabla II.

Tabla II
Porcentaje de solapamiento entre bases de datos (artículos)

	WoS	*MEDLINE*	*IME*
WoS		0,59	0,12
MEDLINE	0,72		0,25
IME	0,17	0,29	

Como se puede observar, el mayor grado de solapamiento de documentos se produce entre MEDLINE y WoS, estando el 72% de los documentos de MEDLINE recogidos también en WoS. Por su parte, el 59% de los documentos de WoS están también en MEDLINE. El menor solapamiento se produce entre IME y WoS y afecta al 17% de los documentos IME y al 12% de los documentos WoS.

4.2. Consecuencias de la inclusión en MEDLINE de un solo lugar de trabajo

La recuperación de documentos de MEDLINE a través de una búsqueda efectuada en el campo lugar de trabajo tiene el inconveniente de que esta base de datos sólo recoge el centro de trabajo del primer firmante, por lo que hay que asumir la pérdida de aquellos documentos en los que la región estudiada firma en una posición diferente a la primera. En este apartado se pretende realizar una estimación de la importancia de dicha pérdida de documentos.

a) Recuperación incompleta de la producción

Una primera aproximación a la pérdida de documentos en MEDLINE consiste en cuantificar el número de artículos en los que la región estudiada firma en una posición diferente a la primera en la base de datos WoS, y extrapolar este comportamiento a la base de datos MEDLINE. La región firma el 33% de sus artículos en WoS en una posición diferente a la primera, por lo que en una primera estimación, su producción MEDLINE estaría infravalorada en un 33%.

Una aproximación más rigurosa se obtiene identificando las revistas comunes a WoS y MEDLINE y cuantificando los documentos no recuperados de MEDLINE a pesar de aparecer en revistas comunes. Una vez desechadas las pérdidas por diferencias en cobertura de tipos documentales u otras razones, se identificaron 625 artículos en los que la región analizada aparecía en segunda posición o posteriores. Esta cifra corresponde al 30% del total de los documentos WoS-MEDLINE, y está muy próxima al 33% antes estimado.

Estos datos indican que el conjunto total de artículos comunes a WoS y MEDLINE ascendería en realidad a 1666 artículos (1.041+625) y que el solapamiento entre WoS y MEDLINE sería superior al observado directamente de los artículos descargados (tabla III). De este modo se puede observar que el 95% de los artículos de WoS están recogidos en MEDLINE, aunque parte de ellos no se pueden obtener de una búsqueda directa por lugar de trabajo a causa de la mencionada limitación de MEDLINE. La tabla III incluye los datos de solapamiento entre WoS y MEDLINE. El solapamiento «observado» se basa en los documentos recuperados mediante las estrategias de búsqueda empleadas, mientras que el «estimado» calcula el solapamiento real entre bases de datos una vez corregido para los defectos en la recuperación de documentos.

Tabla III
Solapamiento observado y estimado entre WoS y MEDLINE (IME no considerado)

	Únicos WoS	*WoS y MEDLINE*	*Total MEDLINE*	*Total WoS*	*Solapamiento WoS-MEDLINE*	*Solapamiento MEDLINE-WoS*
Observado	712 (659+53)	1.041 (155+886)	1.455	1.753	1.041/1.753=0,59	1.041/1.455=0,72
Estimado	87 (712-625)	1.666 (1.041+625)	2.080 (1.455+625)	1.753	1.666/1.753=0,95	1.666/2.080=0,80

Nota: 625 artículos comunes a WoS y MEDLINE no se recuperan de MEDLINE por figurar la región estudiada en un lugar de trabajo diferente al primero.

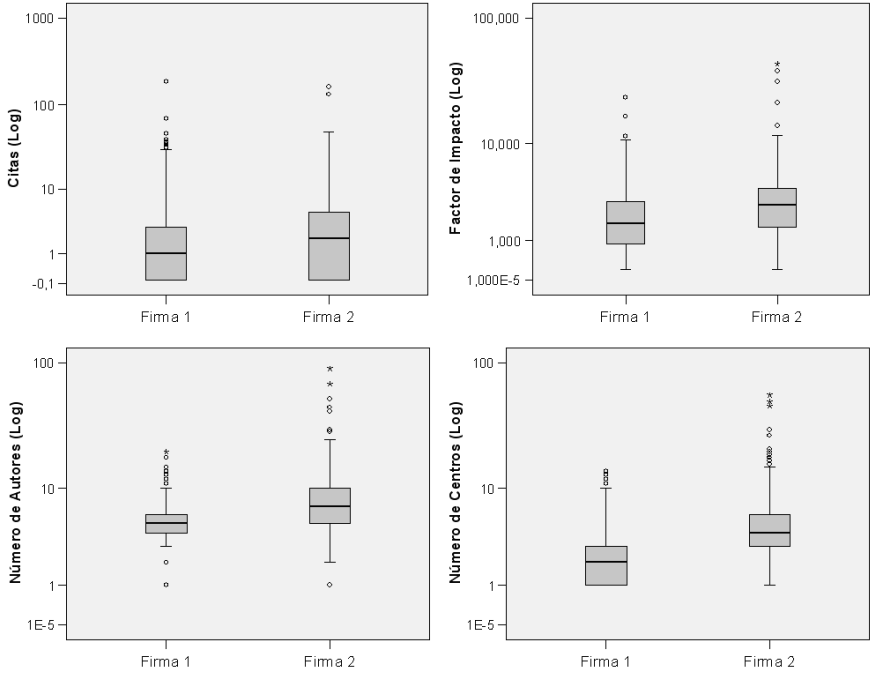
Se pone así de manifiesto la limitación que supone para la realización de estudios bibliométricos delimitados geográficamente la inclusión en MEDLINE de un solo lugar de trabajo. En el caso concreto de la región estudiada, cerca de 1/3 de sus artículos no se recuperarían de la base de datos por figurar dicha región en un lugar de firma distinto al primero. Esta pérdida se produciría en los estudios basados sólo en MEDLINE, pero se reduce si se usan además otras bases de datos con las que existe solapamiento de documentos y que incluyen todos los lugares de trabajo de los autores. Así, el uso combinado de MEDLINE y WoS reduce la pérdida de artículos a un 9%, considerando que del total de los artículos de MEDLINE, el 72% están solapados con WoS y el 28% restante son únicos (0,33*0,28=0,09). Si además se añade IME, el porcentaje de artículos únicos a MEDLINE se sitúa en un 14% y la pérdida se reduciría a un 5% (0,33*0,14=0,05).

b) Impacto

La inclusión de un solo lugar de trabajo en los registros de MEDLINE también puede tener repercusiones sobre el impacto medio calculado para la región sobre dicha producción. Considerando los documentos publicados en revistas comunes a WoS y MEDLINE, se distinguen dos grupos de documentos que vamos a denominar: «Firma 1», formado por los artículos en los que la región firma en primer lugar (identificados en MEDLINE a través de la estrategia de búsqueda empleada); y «Firma 2», que incluye el conjunto de artículos en los que la región firma en segundo lugar o posteriores posiciones (no identificables en MEDLINE a través del lugar de trabajo). Se observa entonces que el grupo «Firma 2» incluye con más frecuencia investigación realizada en colaboración (presenta un mayor numero de autores y centros/documento) ($p<0,000$), que tiende además a publicarse en revistas de más prestigio atendiendo a su factor de impacto ($p<0,000$) y a recibir un mayor número

de citas (p<0,000) que los artículos del grupo «Firma 1». Así, los indicadores de impacto calculados sin tener en cuenta los registros «perdidos» estarían infravalorando la visibilidad e impacto real de la producción de la región (figura 3).

Figura 3
Diferencias en colaboración e impacto de los artículos solapados entre WoS y MEDLINE en los que a) la región firma en primer lugar (Firma 1) (N=1041); b) la región firma en posiciones posteriores (Firma 2) (N=625)



5. Conclusiones

Conocer las ventajas y limitaciones de las distintas bases de datos es esencial para seleccionar la más adecuada para cada estudio, sobre todo si consideramos que dicha selección puede incidir en los resultados de los análisis. Este trabajo muestra que, aunque MEDLINE es la base de datos de referencia en el área biomédica, la inclusión de un solo lugar de trabajo en sus registros constituye un grave inconveniente que limita su uso en los estudios bibliométricos. Hay que señalar que los resultados obtenidos en este trabajo se refieren al comportamiento de una determinada región, y que podrían existir diferencias inter-regionales que habría que analizar en un futuro.

Cobertura, solapamiento y singularidad

A pesar de la amplia cobertura de las ciencias biomédicas repetidamente descrita para la base de datos MEDLINE, ésta solo aporta el 41% de la producción de la región estudiada, mientras que WoS permite recuperar el 65% e IME el 37%. Se observa, pues, que a través de una búsqueda en el campo lugar de trabajo, WoS es la base de datos que permite una mayor recuperación de las publicaciones de la región. El mayor solapamiento se produce entre MEDLINE y WoS, mientras que IME presenta la mayor singularidad, es decir, que cuenta con el mayor porcentaje de documentos únicos no incluidos en las otras bases de datos –todas ellas revistas españolas-, y tiene especial valor para analizar la investigación biomédica con una orientación más local. La baja cobertura y singularidad de MEDLINE se asocia a la imposibilidad de recuperar aquellos documentos en los que la región firma en segunda o posteriores posiciones.

¿Hasta qué punto estas diferencias en cobertura, descritas para una región concreta, son extrapolables al resto de las regiones españolas? Creemos que podrían existir pequeñas diferencias entre regiones en función del peso de la investigación básica y clínica en las mismas. Una mayor orientación básica se acompañaría de un mayor porcentaje de publicaciones internacionales (mejor cobertura en WoS); mientras que el peso de MEDLINE e IME podría incrementarse paralelamente con la actividad clínica. Estudios previos han señalado la mejor cobertura de la base de datos SCI frente a MEDLINE en ciencias biomédicas, mientras que la situación se invierte para la investigación clínica (Pestaña, 1997).

Un solo lugar de trabajo en MEDLINE: implicaciones sobre indicadores de actividad e impacto

Nuestro análisis muestra que si se realizara el estudio de esta región basado solo en MEDLINE se infravaloraría su actividad en cerca de un 30%, ya que éste es el porcentaje de su producción no identificable por el campo lugar de trabajo, porque la base de datos solo incluye la dirección del primer firmante de los documentos. En realidad, MEDLINE proporcionaría más documentos que WoS si incluyera el lugar de trabajo de todos los autores firmantes de un documento. Atendiendo a nuestros datos, las distintas comunidades autónomas españolas firman en segundo lugar o posteriores alrededor del 32% de sus publicaciones en el Web of Science (rango: 26%-46%) (datos de elaboración propia), por lo que podrían producirse pérdidas similares en el caso de otras regiones.

El análisis del conjunto de artículos solapados entre WoS y MEDLINE permite evidenciar el mayor impacto y número de citas recibidas por los artículos de la región firmados en segunda o posteriores posiciones (no identificables a través del campo lugar de trabajo en MEDLINE) en comparación con los firmados en primera posición. Aunque este hallazgo hay que interpretarlo con cautela, dadas las diferencias que existen entre algunas disciplinas en sus hábitos de citación, creemos que

puede ser significativo. En primer lugar, la presencia de colaboración en todos los documentos en los que la región firma en una posición distinta a la primera –todos ellos con al menos dos centros firmantes– puede facilitar el mayor impacto de estos documentos, ya que la colaboración se ha asociado en la literatura a trabajos de mayor prestigio y calidad (véase, por ejemplo, Persson et al., 2004). Los beneficios del trabajo colectivo, en el que se comparten recursos económicos, materiales e intelectuales serían la razón subyacente (Lee y Bozeman, 2005). Pero el mayor índice de citación podría deberse también a una mayor autocitación por parte de los diversos grupos colaboradores (Herbertz, 1995). En cualquier caso, estos artículos no sólo reciben más citas sino que además consiguen situarse en revistas de mayor prestigio, que cuentan con filtros de calidad más estrictos, lo que apunta a que realmente estamos ante una investigación de mayor calidad. Este hecho tiene una implicación importante de cara al desarrollo de estudios bibliométricos: en aquellos basados sólo en MEDLINE no sólo se infravalora cuantitativamente la producción de la región, sino también el impacto medio de la misma.

Complementariedad entre bases de datos

La inclusión de un solo lugar de trabajo en MEDLINE constituye un importante inconveniente para el uso de la base de datos con fines bibliométricos. Aunque el centro primer firmante de una publicación suele corresponder al de los autores más implicados en su desarrollo (Shapiro et al., 1994), el papel de los otros centros firmantes no puede ignorarse, sobre todo ante el creciente papel de la colaboración en la investigación.

Aunque el uso de MEDLINE de forma aislada tiene los mencionados inconvenientes, su uso combinado con otras bases de datos puede ser muy enriquecedor, si se aprovechan las ventajas propias de cada fuente, pudiéndose obtener: a) amplia cobertura: en nuestro caso concreto WoS aporta el 65% de los documentos, la inclusión de IME añadiría un 30% de documentos, y MEDLINE un 6% adicional; b) análisis de contenido: la gran fortaleza de MEDLINE radica en la posibilidad de analizar temas de investigación a través de sus descriptores normalizados, que pueden asignarse a todos los registros solapados entre bases; c) análisis de impacto: a través de diversos indicadores de impacto ligados a WoS e IME; d) análisis de colaboración: a través de los lugares de trabajo incluidos en WoS e IME. Aquí se comentan las tres bases de datos utilizadas en este estudio, pero existen actualmente otras fuentes cuya aportación sería interesante analizar (por ej. Scopus). Como contrapartida a la mayor riqueza de información asociada al uso combinado de varias fuentes de información, existen dificultades técnicas como son las derivadas de la necesidad de integrar datos procedentes de diversas fuentes en una sola base de datos o de detectar registros y revistas solapados para identificar potenciales duplicados (Cañedo Andalia, 1999; Cholvy y Moral, 2001; Hood and Wilson, 2003).

En definitiva, la inclusión en MEDLINE de un solo lugar de trabajo constituye un problema en los estudios bibliométricos delimitados geográficamente, ya que no sólo se infravalora la actividad de la unidad analizada sino también su impacto. El uso combinado de distintas bases de datos reduce la incidencia de estas limitaciones.

Agradecimientos

Se agradece la ayuda técnica de diferentes miembros del grupo ACUTE (Análisis Cuantitativo en Ciencia y Tecnología) del Instituto de Estudios Documentales sobre Ciencia y Tecnología (IEDCYT, antes CINDOC) del CSIC. Este trabajo deriva de un estudio previo para el Instituto Aragonés de Ciencias de la Salud, al que se agradece su interés y apoyo.

Anexo I
Relación de disciplinas incluidas en la delimitación temática de las Ciencias de la Salud en el Web of Science

Alergia, Anatomía y Morfología, Andrología, Anestesiología, Biofísica, Biología Celular, Biología de la Evolución, Biología del Desarrollo, Biométodos, Bioquímica y Biología Molecular, Ciencias del Comportamiento, Cirugía, Corazón y Sistema Cardiovascular, Dermatología, Drogodependencias, Endocrinología y Metabolismo, Enfermedades Infecciosas, Enfermedades Vasculares Periféricas, Enfermería, Ergonomía, Ética Médica, Farmacología y Farmacia, Fisiología, Gastroenterología y Hepatología, Genética y Herencia, Geriatría, Gerontología, Hematología, Inmunología, Medicina Alternativa, Medicina de Urgencia, Medicina Deportiva, Medicina Forense, Medicina Intensiva, Medicina Interna y General, Medicina/Investigación, Medicina Tropical, Medicina/Técnicas de Laboratorio, Micología, Microbiología, Neumología, Neurociencias, Neuroimagen, Neurología Clínica, Nutrición y Dietética, Obstetricia y Ginecología, Odontología y Estomatología, Oftalmología, Oncología, Otorrinolaringología, Parasitología, Patología, Pediatría, Psiquiatría, Química Médica, Radiología y Medicina Nuclear, Rehabilitación, Reproducción, Reumatología, Salud Pública, Medioambiental y Laboral, Servicios Médicos, Toxicología, Trasplantes, Traumatología y Ortopedia, Urología y Nefrología, Virología.

Bibliografía

BEARMAN, T.C.; KUNBERGER, W.A. (1977). *A study of coverage overlap among fourteen major science and technology abstracting and indexing services.* Philadelphia: National Federation of Abstracting and Indexing Services.

BORDONS, M.; GARCÍA-JOVER, F.; BARRIGÓN, S. (1993). Is collaboration improving research visibility? Spanish scientific output in pharmacology and pharmacy. *Research Evaluation,* 3 (1): 19-24

CAÑEDO ANDALIA, R. (1999). Estudios de solapamiento en la selección de las publicaciones seriadas y las bases de datos. *ACIMED,* 7(3): 164-170.

CHOLVY, L.; MORAL, S. (2001). Merging databases: problems and examples. *International Journal of Intelligent Systems,* 16: 1193-1221.

COSTAS, R.; BORDONS, M. (2007). Algoritmos para solventar la falta de normalización de nombres de autor en los estudios bibliométricos. *Investigación Bibliotecológica,* 21 (42): 13-32.

COSTAS, R.; IRIBARREN-MAESTRO, I. (2007). Variations in content and format of ISI databases in their different versions: The case of the Science Citation Index in CD-ROM and the Web of Science. *Scientometrics,* 72 (2): 167-183.

GÁLVEZ, C.; MOYA-ANEGÓN, F. (2007). Standardizing formats of corporate sources data. *Scientometrics,* 70 (1): 3-26.

GLUCK, M. (1990). A Review of Journal Coverage Overlap with an Extension to the Definition of Overlap. *Journal of the American Society for Information Science,* 41 (1): 43-60.

GÓMEZ, I.; GALBÁN, C. (1986). Lack of standardisation in the corporate source field of different databases. *10th International Online Information Meeting.* London: Learned Information.

GÓMEZ, I.; FERNÁNDEZ, M.T.; BORDONS, M.; MORILLO, F.; CANDELARIO, A.; COSTAS, R.; DE FILIPPO, D.; MORENO, L. (2007b). *Producción científica en Ciencias de la Salud de la Comunidad de Aragón y el Servicio Aragonés de Salud (2001-2005).* Madrid: CINDOC-CSIC.

HERBERTZ, H. (1995). Does it pay to cooperate? A bibliometric case study in molecular biology. *Scientometrics,* 33 (1): pp. 117-122.

HOOD, W.W.; WILSON, C.S. (2003a). Informetric studies using databases: Opportunities and challenges. *Scientometrics,* 58 (3): 587-608.

HOOD, W.W.; WILSON, C.S. (2003b). Overlap in bibliographic databases. *Journal of the American Society for Information Science and Technology,* 54 (12): 1091-1103.

LEE, S.; BOZEMAN, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science,* 35 (5): pp. 673-702.

MEYER, DANIEL E.; MEHLMAN, DAVID W.; REEVES, ELLEN S.; ORIGONI, REGINA B.; EVANS, DELORES; SELLERS, DOUGLAS, W. (1983). Comparison study of overlap among 21 scientific databases in searching pesticide information. *Online Review,* 7 (1): 33-43.

MOED, H.F. (2005). Citation analysis in research evaluation. Dordrecht: Springer.

OSCA LLUCH, J. (1999). El Índice Médico Español. *El Profesional de la Información,* 8(4). <http://www.elprofesionaldelainformacion.com/contenidos/1999/abril/el_indice_medico_espaol.html>. Consultado: 1/9/2007.

PERSSON, O.; GLÄNZEL, W.; DANELL, R. (2004). Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics,* 60 (3): 421-432.

PESTAÑA, A. (1997). El Medline como fuente de información bibliométrica de la producción española en biomedicina y ciencias médicas. Comparación con el Science Citation Index. *Medicina Clínica,* 109: 506-511.

PULGARÍN, A.; ESCALONA, M.A. (2007). Medida del solapamiento en tres bases de datos con información sobre Ingeniería. *Anales de Documentación,* 10: 335-344.

RODRÍGUEZ GAIRÍN, J.M.; SOMOZA FERNÁNDEZ, M. (2003). La gestión del conocimiento a partir de estudios bibliométricos. La producción científica española en Medline-PubMed (1997-2002). 1. El problema de los estudios basados en afiliación. En: *X Jornadas Nacionales de Información y Documentación en Ciencias de la Salud, Málaga (Spain).* http://bd.ub.es/pub/rgairin/publicaciones.php.

RUIZ-PÉREZ, R.; DELGADO LÓPEZ-CÓZAR, D.; JIMÉNEZ CONTRERAS, E. (2002). Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies. *Journal of Medical Library Association.* 90 (4): 411-430.

SHAPIRO, D.W.; WENGER, N.S.; SHAPIRO, M.F. (1994). The contributions of authors to multiauthored biomedical research papers. *Journal of the American Medical Association,* 271 (6): 438-442.

SPINAK, E. (1995). Errores ortográficos en el ingreso en Bases de Datos. *Revista Española de Documentación Científica,* 18 (3): 307-309.

TORRES SALINAS, D. (2007). *Diseño de un sistema de información y evaluación científica. Análisis ciencimétrico de la actividad investigadora de la Universidad de Navarra en el área de ciencias de salud. 1999-2005.* Tesis Doctoral. Universidad de Granada.